A Deep 2D Convolutional Network for Waveform-Based Speech Recognition

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Deep 2D Convolutional Network for Waveform-based Speech Recognition

Dino Oglic [1], *Zoran Cvetkovic* [1], *Peter Bell* [2], *and Steve Renals* [2]

[1] Department of Engineering, King's College London, UK
[2] Center for Speech Technology Research, University of Edinburgh, UK
{dino.oglic, zoran.cvetkovic}@kcl.ac.uk, {peter.bell, s.renals}@ed.ac.uk

Abstract

Due to limited computational resources, acoustic models of early automatic speech recognition (ASR) systems were built in low-dimensional feature spaces that incur considerable information loss at the outset of the process. Several comparative studies of automatic and human speech recognition suggest that this information loss can adversely affect the robustness of ASR systems. To mitigate that and allow for learning of robust models, we propose a deep 2D convolutional network in the waveform domain. The first layer of the network decomposes waveforms into frequency sub-bands, thereby representing them in a structured high-dimensional space. This is achieved by means of a parametric convolutional block defined via cosine modulations of compactly supported windows. The next layer embeds the waveform in an even higher-dimensional space of high-resolution spectro-temporal patterns, implemented via a 2D convolutional block. This is followed by a gradual compression phase that selects most relevant spectro-temporal patterns using wide-pass 2D filtering. Our results show that the approach significantly outperforms alternative waveform-based models on both noisy and spontaneous conversational speech (24% and 11% relative error reduction, respectively). Moreover, this study provides empirical evidence that learning directly from the waveform domain could be more effective than learning using hand-crafted features.
Index Terms: automatic speech recognition, parametric filters, deep convolutional networks, raw speech, robustness.

1. Introduction

Scalable and effective acoustic models for speech recognition are typically based on hand-crafted features designed according to the physiology of human hearing and psychoacoustic measurements [1, 2, 3]. The most effective and widely used feature extraction techniques employ band-pass filtering of signals such as log Mel-filter bank values (FBANK) [4] and their decorrelated variant known as Mel frequency cepstral coefficients (MFCC) [1, 5]. A potential shortcoming of these approaches is the fact that the parameters specifying such a representation of a raw speech frame are fixed a priori and not learned using the available data. As a result, feature extraction might be discarding information relevant to robustness, and moreover, is done independently of model learning and it does not necessarily provide an ideal inductive bias for the learning process.

An alternative to learning a discriminative model with statically extracted features is to learn these features automatically as part of a neural architecture that takes raw speech as input. In addition to having a more flexible inductive bias such a model would be less susceptible to the information loss that is inherent to waveform compression by means of a projection to a lower dimensional feature space [6, 7]. In particular, a model operating directly in the waveform domain has a potential to exploit local correlations within the signal that are typically discarded when computing Mel-filter bank values [8], as well as the information contained in a sequence of waveform samples without interruptions by frame boundaries characteristic of spectrograms and non-adaptive feature extraction techniques based on frame-based discrete Fourier transforms [9]. As a result of the latter, phonetic events on the boundaries of short frames are typically poorly described by filterbank features. While there are many benefits of operating directly in the waveform domain there are also some challenges in extracting the information from these high dimensional and highly correlated inputs. In particular, one of the issues recognized in early acoustic models based on raw waveforms is that for a given phonetic unit such inputs are characterized by a large number of variations in the form of phase shifts and temporal distortions [2, 10]. Thus, an effective neural architecture needs to be able to automatically extract features that are invariant to small phase shifts and distortions. Another difficulty in operating with speech waveforms is the high dimensionality of the input space, which requires a large number of parameters [9] and prolonged training time.

A desirable property of an effective representation is invariance to nuisance transformations such as translations [11] and stability to actions of small diffeomorphisms that distort/warp signals [12, 13]. To learn a representation robust to such perturbations of a signal is arguably one of the most important unresolved problems in speech recognition. The empirical effectiveness of state-of-the-art convolutional neural networks can be to a large extent attributed to their ability to encode invariance to local translations via convolutional weight sharing and pooling operators [11, 14]. More specifically, due to their local connectivity patterns convolutional layers are well suited to model local correlations, as well as translations in spectro-temporal waveform decompositions, that can occur as a result of different speaking styles, variations between speakers, additive noise, channel degradation, etc. This type of inductive bias has also been used previously to achieve phase invariance in waveform-based models [8, 10, 15].

We propose a deep 2D convolutional architecture for learning an effective acoustic model directly in the waveform domain. The main idea is to first increase the dimension of the instance space in a structured manner, embedding redundancies into the waveform representation such that it could withstand a significant amount of additive noise and distortion without significant overlaps between different phonetic units [7, 16, 17]. To expand the information present in a waveform signal and allow more flexible feature extraction, we rely on a family of band-pass filters (Section 2) that are defined via cosine modulations of compactly supported Parzen windows. This is a parametric convolutional block that splits a waveform frame into frequency sub-bands and embeds it into a high dimensional but structured space. The dimension of the embedding is further increased (e.g., by a factor of 200 compared to the input frame) by means of a non-parametric 2D convolutional layer. This is followed

http://dx.doi.org/10.21437/Interspeech.2020-1870

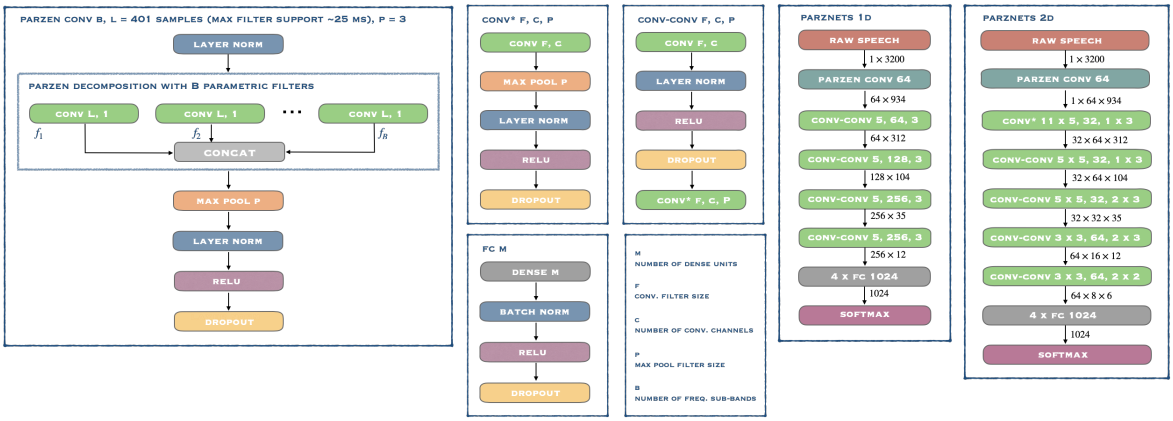

Figure 1: *The figure describes the architectures for* PARZNETS *with* 1D *and* 2D *convolutional operators. This is supplemented with an illustration of Parzen convolutional block (the leftmost panel) that decomposes a raw speech frame into frequency sub-bands.*

by a gradual compression phase that extracts a low-resolution spectro-temporal decomposition by means of standard wide-pass 2D convolutional filtering. In our empirical analysis (Section 3), we show that the network consistently outperforms feedforward models based on non-adaptive feature extraction techniques, as well as state-of-the-art models based on raw speech. This is done on speech recognition benchmark datasets having different properties. In particular, the network: *i*) does not overfit and outperforms all the feedforward architectures on a small TIMIT dataset, *ii*) learns a noise robust representation on AURORA4 and outperforms state-of-the-art very deep convolutional networks for statically extracted features [18, 19], *iii*) outperforms recently proposed architectures for raw speech [20, 21] and performs on par with a state-of-the-art FBANK/MFCC based TDNN [2] model on AMI (conversational speech, without i-vectors and data augmentation). Thus, the empirical contributions provide the first comprehensive evidence for the effectiveness of learning artificial neural networks directly from waveform, as opposed to building them on top of statically extracted features.

2. Parznets

In speech recognition, band-pass filtering of signals is traditionally performed by weighted averaging of power spectra [5, 13], computed over segments of fixed duration. Alternatively, the signal can be convolved by a filter directly in the time-domain. Motivated by this, we design the first layer in our architecture to emulate this operation via a parametric time-domain convolutional operator. To that end, we consider a family of differentiable band-pass filters based on cosine modulations of compactly supported Parzen windows [22]. In particular, our empirical analysis employs squared Epanechnikov window function [23]

$$k_\gamma(t) = \begin{cases} \left(1-\gamma t^2\right)^2 & |t| \leq 1/\sqrt{\gamma} \\ 0 & \text{otherwise} \end{cases},$$

where γ is a parameter controlling the window width. To allow for flexible placement of the center/modulation frequency, we rely on cosine modulation. Thus, Parzen filters are defined with only two differentiable parameters, η controlling the modulation frequency and γ controlling the filter bandwidth, i.e.,

$$\phi_{\eta,\gamma}(t) = \cos(2\pi\eta t) \cdot k_\gamma(t) \ . \quad (1)$$

As the filters are real-valued, the corresponding convolutions are simpler to implement compared to their complex-valued counterparts with exponentially modulated windows [24]. As illustrated in Figure 1 (the leftmost panel), for each filter configuration $\{(\eta_i, \gamma_i)\}_{i=1}^{B}$, we use Eq. (1) to generate a one dimensional filter with maximum length given by the number of samples in 25 ms of speech; filters with shorter support are symmetrically padded with zeros. In comparison to wavelet filters [25], the Parzen convolutional block offers additional flexibility by allowing independent control over bandwidth and modulation frequency. The outputs of parametric convolutions are concatenated into a spectro-temporal decomposition of a signal and then passed to a max pooling operator, followed by layer normalization [26].

The motivation behind the Parzen block is to embed the signal into a structured high dimensional space where, we hypothesize, phonetic units will be easier to separate. Moreover, by increasing the dimension of the space further via a 2D non-parametric convolutional layer (Figure 1, PARZNETS 2D, CONV*) we aim to embed redundancies into the representation such that it could withstand signal corruption, while still keeping separability between phonetic units. At the output of such a CONV* block the dimension of the embedding is increased by a factor of 200 compared to the input frame size. The outputs of that non-parametric embedding block are then passed to a sequence of double convolutional blocks that perform further band-pass filtering and compression of the signal by different max pooling operators. The convolutional blocks generate a set of *automatically extracted features*, which are then passed to a multi-layer perceptron with 4 hidden layers. We rely on the RELU non-linearity throughout the network as it has recently been established that such networks can be made robust under ℓ_p perturbations [27].

The main challenge for a neural architecture with 2D convolutions (Figure 1, PARZNETS 2D) that takes high dimensional input, is to design an effective compression operator. The dimensions of the time and frequency axes of the sub-band decomposition differ significantly and one cannot take identical compression factors across them. The rationale behind the initial convolutional filter size over time comes from dynamic DELTA and DELTA-DELTA features [28], typically combined with FBANK and MFCC coefficients, that are essentially realized by applying 5-tap wide convolutional operators. We initially adhere to that filter configuration over the time-domain and employ the 11-tap wide filter over frequency bands in the first 2D convolutional layer, i.e., filter size 11×5. Following this, we switch to a compression regime using the double convolution block with 5×5 filters (Figure 1, CONV-CONV), combined with max pooling compression 1×3, that retains all frequency components. At this stage the number of frequency bands and time samples

are approximately of the same scale and we combine another block of double convolutions with max pooling compression operator of size 2×3, starting to compress over the frequency domain, too. The resulting spectro-temporal decomposition is of low resolution and we can employ convolutions with filter size 3×3, known for performing well in computer vision and speech recognition at such resolutions [18]. After the first such double convolution block we compress with a factor of 2×3 and then finally after another such block with a factor of 2×2. The resulting *automatically extracted features* are then passed to a multi-layer perceptron with 4 hidden layers.

In addition to evaluating PARZNETS 2D neural architecture relative to state-of-the-art baselines for waveform-based speech recognition, we also consider its merits relative to a convolutional architecture based on 1D convolutions (Figure 1, PARZNETS 1D).

3. Experiments

We evaluate PARZNETS on three different benchmark datasets: TIMIT [29], AURORA4 [30], and AMI [31]. The goal of the first experiment on TIMIT is to demonstrate that architectures based on raw speech, such as PARZNETS do not require large training datasets to outperform models based on non-adaptive features. In the second experiment on AURORA4, we aim to show that PARZNETS can learn a noise robust representation of waveform signals. In the third experiment on AMI, we demonstrate that PARZNETS generalize to learning from conversational speech and outperform state-of-the-art raw waveform based approaches.

In all of our experiments, we train a context dependent model based on frame labels (i.e., HMM state ids) generated using a triphone model from Kaldi [32] with 25 ms frames and 10 ms stride between the successive frames. The data splits (training/development/evaluation) are identical to the ones from the corresponding Kaldi recipes. In the preprocessing step, we assign the Kaldi frame label to a 200 ms long segment of raw speech centered at the original Kaldi frame. The Parzen convolution block is initialized by taking the modulation frequencies to be equidistant in mel-scale. The bands of filters are initialized as in FBANK features. For convolutional and dense blocks in our network, we employ the Xavier initialization scheme [33] with magnitude 0.005. While the convolutional blocks are initialized with the factor type *in*, the dense blocks use the *avg* type. The feature extraction layers (i.e., Parzen and convolutional parameters) are updated using the RMSPROP optimizer with initial learning rate set to 0.0008. The multi-layer perceptron blocks are updated using stochastic gradient descent with initial learning rate set to 0.08. A similar combination of optimizers (all network parameters are optimized jointly) was used in [21]. After the relative validation error falls below 0.1%, we decrease the learning rates by a factor of 2. We use the minibatch size of 512 samples and terminate the training process after 25 epochs.

3.1. TIMIT

To be consistent with our baselines (neural architectures for raw speech) on TIMIT, we generate frame labels (1 912 HMM state ids) using the DNN triphone model and decoding configuration from [21]. Table 1 summarizes our results relative to state-of-the-art feedforward architectures on this relatively small dataset. A comparison to previously reported results for raw speech baselines shows that our PARZNETS 2D architecture with two dimensional convolutions performs the best on average and appears not to overfit on this small dataset, despite being a rather deep architecture. Moreover, this is the first neural architecture for raw speech that outperforms significantly feedforward models paired with standard statically extracted features. Note that lower phone error rates have been observed on TIMIT using recurrent networks [37] (LI-GRU: 15.8%; LI-GRU-FMLLR: 14.8%), and using 960 hours of LIBRISPEECH for unsupervised pretraining [38] (VQ-WAV2VEC+BERT: 11.4%).

Table 1: *The phoneme error rates obtained on the test set of* TIMIT *with various input features and neural architectures.*

ARCHITECTURE	AVG	MIN
A. RAW SPEECH		
PARZNETS 1D	17.2	17.1
PARZNETS 2D	**16.6**	**16.3**
SINCNET [21, 34]	17.5	17.2
SINC^{2}NET [35]	-	16.9
RAW SPEECH CNN [34]	18.3	18.1
END-TO-END CNN [24]	-	18.0
B. STANDARD FEATURES		
MFCC-MLP	18.1	17.8
FMLLR-MLP	16.9	16.7
M-DSS I & II + CNN & MLP [36]	-	17.4

3.2. AURORA4

AURORA4 is a standard benchmark for noisy speech with signal corruptions due to convolutional and additive noise, different microphones, and the mismatch between training and test samples. We focus here on multi-condition training and show that the proposed architecture outperforms all previously evaluated feedforward architectures, irrespective of the input domain (raw speech or standard features). To be consistent with the baselines, we generate alignments using both GMM and DNN triphone models (3 408 and 2 016 HMM state ids, respectively). Table 2 summarizes our results relative to relevant baselines on this dataset. In comparison to state-of-the-art convolutional model [18] based on non-adaptive FBANK features (VDCNN with two dimensional convolutions), our approach does statistically significantly better (the Wilcoxon test with 95% confidence). Recently, a novel type of multi-octave convolution [19] has been proposed for FBANK features and our empirical results show that PARZNETS 2D with simple two dimensional convolutions performs on par with that much more complex architecture. We also compare to the SINCNET architecture (state-of-the-art for raw speech) and our results demonstrate that we statistically significantly outperform this approach using both GMM and DNN alignments. Moreover, the considered PARZNETS architectures also outperform multi-layer perceptrons (MLP) with FMLLR and MFCC features.

3.3. AMI

AMI-IHM is a conversational speech dataset with approximately 78 hours of speech, recorded using individual headset microphones. We generated alignments using the Kaldi recipe configured with 3 984 HMM state ids. Table 3 summarizes our result relative to relevant baselines on this dataset.

We compared PARZNETS with two recently published raw waveform approaches for this task: multi-span raw waveform models [20] and SINCNET [39], and show that PARZNETS obtains over 10% relative improvement in WER compared to these methods. Moreover, we also compare to deep time-delay networks [40] based on FBANK/MFCC features (considered to be state-of-the-art feedforward model on this dataset) and show that the proposed architecture with two dimensional convolutions performs on par with that approach. We note here that we have not used any data augmentation or i-vectors in our experiments, both techniques which could be used with our approach. More-

Table 2: *The word error rates (%) obtained on different test sets of* AURORA4 *with various input features and neural architectures.*

TEST SET	DNN ALIGNMENTS			GMM ALIGNMENTS						
	RAW SPEECH			RAW SPEECH			STANDARD FEATURES			
	PARZNETS 1D	PARZNETS 2D	SINCNET	PARZNETS 1D	PARZNETS 2D	SINCNET	MFCC-MLP	FMLLR-MLP	VD10CNN2D [18]	M-OCT CNN [19]
A_1	2.52	2.32	3.12	2.80	3.01	3.42	4.28	3.34	3.27	2.32
B_{2-7}	4.61	4.38	5.97	4.80	4.74	6.33	7.44	6.27	5.61	4.73
C_8	5.06	4.30	5.68	5.14	4.99	6.13	8.73	5.74	5.32	4.24
D_{9-14}	14.78	12.73	16.58	14.41	13.15	16.99	18.71	16.04	13.52	13.57
AVG_{1-14}	8.85	**7.80**	10.29	8.80	**8.24**	10.68	12.14	10.21	8.81	8.31

over, GPU memory limitations meant that our experiments were performed with a modest number of channels applied to the high dimensional raw waveform inputs (see Figure 1) and we anticipate that the results could be further improved with increased number of parameters (i.e., CONV channels). Finally we note that our experiments were conducted using a cross entropy (CE) loss function. Experiments using a sequence discriminative approach (LF-MMI) indicate that the WERs could be further lowered – Povey et al [41] indicated that using LF-MMI in place of CE can reduce the error rate by about 10% relative on this task, and more recently a regularised LF-MMI training with significant data augmentation (6x) resulted in a WER of 18.0% on this task [42]. Our future work will explore sequence discriminative training for PARZNETS.

Table 3: *The word error rates obtained on dev and eval/test sets of* AMI-IHM *with various input features and neural architectures.*

ARCHITECTURE	DEV	EVAL
A. RAW SPEECH		
PARZNETS 1D	25.5	26.6
PARZNETS 2D	**24.9**	**26.0**
SINCNET [39]	28.0	30.2
MULTI-SPAN-DNN [20]	27.2	29.3
B. STANDARD FEATURES		
FBANK-MLP [20]	28.3	31.1
FMLLR-MLP	26.0	27.1
TDNN [40]	25.3	**26.0**

4. Discussion

In previous work on raw waveform based speech recognition, it has been observed that such models can outperform non-adaptive feature extraction techniques in the multi-microphone setting [8, 10, 43]. Another common finding was that in the single microphone setting with more than 2 000 hours of training data, neural architectures with raw speech inputs are on par or sometimes even better than models based on FBANK, MFCC, or FMLLR features. To the best of our knowledge, there has not been a comprehensive empirical study showing that learning directly from the waveform domain can be more effective than learning with statically extracted features across different environments (small training datasets, noisy data, mismatch between train and test sets, spontaneous conversational speech).

In the majority of previously considered architectures there is typically a single convolutional layer with 1D convolutions, designed to emulate log filterbank magnitude features [8, 15]. Sainath et al. [10] propose an architecture which takes raw waveform inputs and applies time-domain followed by frequency-domain one dimensional convolutions, designed to extract band-pass features from the waveform. The extracted features are then passed to a sequence of long short-term memory (LSTM) blocks that capture the sequential relations between the inputs. The architecture requires more than 2 000 hours of training data to match the performance of neural architectures with non-adaptive features. Similarly, Zhu et al. [44] combine two convolutional layers with recurrent blocks in end-to-end training, requiring more than 2 400 hours of training data for state-of-the-art results. Ghahremani et al. [2] proposed a feedforward architecture based on convolutional feature extraction layer, with the outputs of that block passed to a TDNN. The empirical evidence indicates that the approach is competitive with MFCC-based architectures on large datasets. The model has not been evaluated on noisy speech and it is unclear how well it generalizes on small datasets.

PARZNETS are based on parametric convolutions: perhaps the most prominent related work is the SINCNET architecture [21], which is considered to be the state-of-the-art for raw waveform speech recognition at the moment. The architecture employs three 1D convolutional layers on top of a parametric convolution block. A related architecture is SINC^{2}NET that links a parametric convolution block to an MLP [35]. Recently, complex-valued parametric filters have been used to initialize a complex non-parametric convolution block in a deep convolutional network for end-to-end speech recognition [24, 45, 46]. The architecture relies on 1D convolutions with a large number of channels and requires 1 000 epochs for convergence [24]. In comparison to [24], we demonstrate that our 2D architecture generalizes better on the small TIMIT dataset. For our experiments, we have picked the SINCNET architecture (code available online) as a representative baseline from this class and showed that the proposed architectures outperform it across different datasets.

Recently, an approach based on concatenation of multiple convolutional blocks was proposed [20], in which convolutional blocks capture different contexts in time and learn band-pass filters that are more expressive than classic Mel-filterbanks which operate on a single fixed context. The approach has been evaluated on both noisy and conversational speech. In our experiments, we have compared to this baseline and demonstrated statistically significant improvement on the AMI dataset.

5. Conclusion

We have proposed deep 2D convolutional networks – PARZNETS – for robust speech recognition in the waveform domain and demonstrated generalization across different settings. Our empirical results demonstrate that the PARZNETS 2D architecture consistently outperforms alternative feedforward models on both noisy and conversational speech. To the best of our knowledge, this is the first comprehensive empirical study showing that learning directly from the waveform domain can be more effective than learning using statically extracted band-pass features. A more elaborate analysis of the architecture in terms of selecting a good number of channels in convolutional layers has been hindered by the limited capacity of our GPU devices. However, even the network with modest number of channels across convolutional layers has managed to either outperform or match state-of-the-art feedforward models on the considered datasets.

6. Acknowledgements

This work was supported in part by EPSRC grant EP/R012067/1 (SPEECHWAVE). The authors would also like to thank Erfan Loweimi and Neethu Joy for constructive feedback and help with Kaldi alignments.

7. References

[1] J. S. Bridle and M. Brown, "An experimental automatic word-recognition system," JSRU, Ruislip, UK, Tech. Rep. 1003, 1974.

[2] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INTERSPEECH*, 2016.

[3] Z. Tüske, R. Schlüter, and H. Ney, "Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing," in *IEEE ICASSP*, 2018.

[4] S. Pruzansky, "A pattern-matching procedure for automatic talker recognition," *JASA*, vol. 35, pp. 354–358, 1963.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 28, pp. 357–366, 1980.

[6] M. Ager, Z. Cvetkovic, and P. Sollich, "Combined waveform-cepstral representation for robust speech recognition," *IEEE ISIT*, 2011.

[7] ——, "Speech recognition front end without information loss," *arXiv:1312.6849*, 2015.

[8] Y. Hoshen, R. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *IEEE ICASSP*, 2015.

[9] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, 2014, pp. 890–894.

[10] T. Sainath, R. J. Weiss, K. Wilson, A. W. Senior, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, 2015.

[11] A. Raj, A. Kumar, Y. Mroueh, T. Fletcher, and B. Schölkopf, "Local group invariant representations via orbit embeddings," in *AISTATS*, 2017.

[12] A. Trouvé and L. Younes, "Local geometry of deformable templates," *SIAM Journal on Mathematical Analysis*, 2005.

[13] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, 2012.

[14] R. Kondor and S. Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups," in *ICML*, 2018, pp. 2747–2755.

[15] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *INTERSPEECH*, 2013.

[16] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, pp. 1396–1407, 2011.

[17] J. Yousafzai, P. Sollich, and Z. Cvetkovic, "Noise robust phoneme classification in frequency subbands using support vector machines," *IEEE SLT*, 2010.

[18] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Trans. on Audio, Speech and Lang. Proc.*, 2016.

[19] J. Rownicka, P. Bell, and S. Renals, "Multi-scale octave convolutions for robust speech recognition," in *IEEE ASRU*, 2019.

[20] P. von Platen, C. Zhang, and P. Woodland, "Multi-span acoustic modelling using raw waveform signals," in *INTERSPEECH*, 2019, pp. 1393–1397.

[21] M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with SincNet," *arXiv:1812.05920*, 2018.

[22] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, 1962.

[23] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, vol. 14, pp. 153–158.

[24] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schatz, G. Synnaeve, and E. Dupoux, "Learning filterbanks from raw speech for phone recognition," in *IEEE ICASSP*, 2018.

[25] H. Khan and B. Yener, "Learning filter widths of spectral decompositions with wavelets," in *NeurIPS*, 2018.

[26] L. J. Ba, R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[27] F. Croce and M. Hein, "Provable robustness against all adversarial l_p-perturbations for $p \geq 1$," in *ICLR*, 2020.

[28] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 34, pp. 52–59, 1986.

[29] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. of DARPA Workshop on Speech Recognition*, 1986.

[30] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," MSU ISIP, Tech. Rep., 2002.

[31] S. Renals, T. Hain, and H. Bourlard, "Recognition and interpretation of meetings: AMI and AMIDA," in *IEEE ASRU*, 2007.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.

[34] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," *arXiv:1811.07453*, 2019.

[35] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[36] V. Peddinti, T. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep scattering spectrum with deep neural networks," in *IEEE ICASSP*, 2014.

[37] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Tran. Emerging Topics in Computational Intelligence*, vol. 2, pp. 92–102, 2018.

[38] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv:1910.05453*, 2019.

[39] J. Fainberg, O. Klejch, E. Loweimi, P. Bell, and S. Renals, "Acoustic model adaptation from raw waveforms with SincNet," in *IEEE ASRU*, 2019.

[40] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[41] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH*, 2016.

[42] N. Kanda, Y. Fujita, and K. Nagamatsu, "Lattice-free state-level minimum Bayes risk training of acoustic models." in *Interspeech*, 2018.

[43] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, B. Li, E. Variani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Raw multichannel processing using deep neural networks," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*. Springer, 2017.

[44] Z. Zhu, J. Engel, and A. Hannun, "Learning multiscale features directly from waveforms," in *INTERSPEECH*, 2016.

[45] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," in *INTERSPEECH*, 2018, pp. 781–785.

[46] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, "Fully convolutional speech recognition," *arXiv:1812.06864*, 2018.